2021 54th Hawaii International Conference on System Sciences

Weather and Random Forest-based Load Profiling Approximation Models and Their Transferability across Climate Zones

Huifen Zhou
Pacific Northwest National Laboratory
huifen.zhou@pnnl.gov

Z. Jason Hou
Pacific Northwest National Laboratory
zhangshuan.hou@pnnl.gov

Yuan Liu
Pacific Northwest National Laboratory
yuan.liu@pnnl.gov

Pavel Etingov
Pacific Northwest National Laboratory
pavel.etingov@pnnl.gov

Abstract

This study is to provide predictive understanding of the associations of weather attributes with electricity load profiles across a variety of climate zones and seasons. Firstly, machine learning (ML) approaches were used to identify and quantify the impacts of various weather attributes on residential and commercial electricity demand and its components across the western United States. Performance and transferability of the developed ML models were then evaluated across different temperate zones (e.g., southern, middle, and northern US) and across coastal, mid-continent, and wet zones, with inputs of weather condition data from the National Oceanic and Atmospheric Administration (NOAA) at representative weather stations. The predictive models were developed based on the ranked and screened factors using the regression tree (RT) and random forest (RF) approaches, for five different scenarios (seasons).

1. Introduction

Load composition varies temporally and spatially across bulk electric system interconnections, posing challenges to power grid modeling and simulation. Accurate estimation of component-wise load shape is pivotal in power system modeling and analysis, and is particularly important when large contingencies take place within a specific timeframe [1, 2]. Defective modeling of load composition could incapacitate the simulation model from tracking the actual power system behaviors [3, 4]. However, load composition estimation is difficult to accomplish because of insufficient data sources and non-uniform load categories.

Efforts have been made to advance the load composition modeling considering different factors, such as different climate zones and weather information. For example, load composition data were updated for the entire WECC system, using only the up-to-date measured load profile data from the Pacific Northwest regions [5], or by evaluating the cross correlation information between old and updated load profiles and applying the correlation coefficient-based weighting factors to calculate the load profiles for the regions without updated load data [6]. A more advanced machine learning (ML)-based approach was also implemented to estimate load composition profiles for regions without effective datasets [7].

The usage of energy is important to economy and power supply companies especially under critical weather conditions [7-9], such as hot summers when the usage of cooling in some southern areas of the United States is significant. It has been found that the variation of temperature affects the heating and cooling usages throughout the year. Wan et al. [10] used the dry ball temperature, wet ball temperature and global solar radiation to analyze how the energy use of the office buildings responded to climate change. Lindberg et al. [11] studied the relationship between building energy consumption and weather attributes, and they found that the heating usage was affected by temperature. Beccali et al. [12] used the weather attributes and electricity intensity data to predict one hour ahead load consumption; They pointed out that humidity index can be used to infer the household electricity consumption.

ML approaches such as tree-based approach, support vector machine (SVM) and artificial neural network (ANN) have been widely applied to learn the impact of weather attributes on electricity load [12-18]. Li et al. [17] applied SVM to predict hourly cooling load in the building by using outdoor dry ball temperature and solar radiation. Mori et al. [18] used a hybrid technique of the optimal regression tree and ANN method to develop short-term load forecasting, in which temperature and humidity are taken as input variables.

In our previous work, based on the load data of Western Electricity Coordinating Council (WECC) and climate conditions in these areas, we studied the relationships between weather conditions and electricity load of commercial and residential customers by using regression tree (RT) and random

URI: https://hdl.handle.net/10125/71018
978-0-9981331-4-0

forest (RF) approaches with systematic cross-validation. However, due to the limitation such as missing or lack of load survey data, it is difficult to approximate the load profiles in the Eastern zones. Based on the WECC climate zonation system and the international energy conservation code climate region definition [19], it is possible to develop transferrable load profile approximation models using only the WECC system data, as the WECC training data covers different temperature and climate zones [7], and one can develop global or zone-specific models depending on the climate conditions.

In this paper we develop and evaluate the transferability of the RF models trained using the WECC data across climate regions with varying temperatures (mainly north-to-south) and humidity (mainly along east-west direction). A reliable transferable RF model enables prediction of electricity usage with only local weather information.

2. Data

We adopt the climate zone definition and delineation based on the International Energy conservation code, as shown in Figure 1. The climate classification system provides more details describing the differences in humidity and temperature variations (see Figure 2).

2.1. Load Data

The load data from 2011-2012 Northwest Energy Efficiency Alliance (NEEA) residential building stock assessment (RBSA) [8] and 2006 California Commercial End-Use Survey (CEUS) [9] are used. The raw data is processed to account for 24-hour variations at hourly resolution, 12 climate zones, and five seasons, along with various end-use types including cooling, lighting, heating, ventilation, and so on.

2.2. Weather Data

Other variables include climate zone, season, hour of the day, and climate zone index. Weather data were collected from the National Oceanic and Atmospheric Administration (NOAA) website, with the same time spans as the load data, at one representative weather station for each climate zone. The weather attributes include six variables: visibility (%), temperature (deg C), dew point temperature (deg C), humidity (%), wind speed (mph), and precipitation (inch), for each climate zone, season, and hour of the day.

2.3. International Energy Conservation Code Climate Regions

International Energy conservation code climate regions map [19, 20] divides the United States into 8 (1 – 8) temperature-based climate zones (1 is the hottest area and 8 is the subarctic area), and 3 humidity-based regimes which are A (moist), B (dry) and C (Marine), as shown in Figure 1. The novel way of dividing the zones longitudinally and laterally and the unique dataset covering the US and with all the load components enables a comprehensive understanding of the associations between load components and weather attributes.

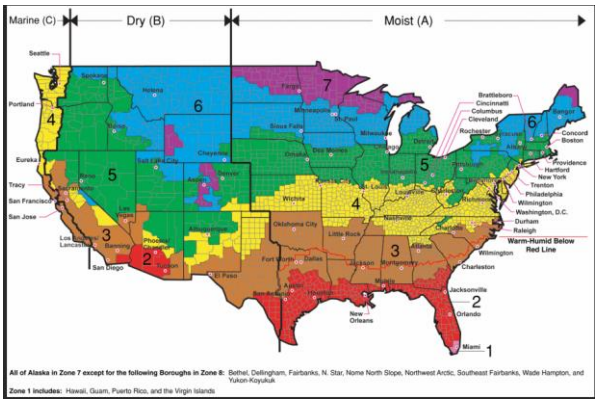


Figure 1. International Energy conservation code climate regions [20].

2.4. Representative Cities for Different Climate Zones

Representative cities were selected, one for each climate region. The corresponding temperature zones and humidity-based climate regimes (i.e., humidity zones) are listed in table 1.

Table 1. Climate regions and representative cities.

Representative Cities	State	Temperature zone	Climate regime
Seattle	WA	4	C
Portland	OR	4	C
Boise	ID	5	B
Billings	MT	6	B
San Francisco	CA	3	C
Sacramento	CA	3	B
Fresno	CA	3	B
San Diego	CA	3	B

Los Angeles	CA	3	C
Bakersfield	CA	3	B
Phoenix	AZ	2	B
Las Vegas	NV	3	B
EL Paso	TX	3	B
Denver	CO	5	B
Reno	NV	5	B
Albuquerque	NM	4	B
Salt Lake City	UT	5	B
Minneapolis	MN	6	A
Wichita	KS	4	A
Dallas	TX	3	A
Houston	TX	2	A
Chicago	IL	5	A
Nashville	TN	4	A
Toronto	ONT	6	A
Boston	MA	5	A
Baltimore	MA	4	A
New Orleans	LA	2	A
Tampa	FL	2	A
Indianapolis	IN	5	A
Pittsburgh	PA	5	A
Oklahoma	OK	3	A
Savannah	GA	2	A
Atlanta	GA	3	A
Charlotte	NC	3	A

3. Methodology

3.1. Regression Tree Method

Breiman et al. [21] introduced the classification and regression tree (CART) approach. Tree-based model split data into multiple unit interval with continuous response variable Y and binary inputs X_1 and X_2. Then the recursive portioning results in multiple region Rm where the model predicts Y by using those multiple regressions [22]:

$$f\left(x\right) = \sum_{m=1}^{M} c_m I\left(x \in R_m\right) \quad (1).$$

In the trees, the regions R_m are usually defined by means of binary split, and I(.) is an indicator function returning 1 if its argument is true and 0 if otherwise, and M is the number of partition regions. For a data set, we would like to pick the regions R_m and the constants cm to minimize the squared error.

$$\sum(f(x_i) - y_i)^2 \quad (2).$$

Then the overall sums of squares error are minimized:

$$minimize\ SSE = \sum_{c=1}^{m} \sum_{i=1}^{C} (y_i - C_m)^2 \quad (3).$$

The size of trees is reduced by removing sections that provide little power (e.g., in terms of mean squared errors) to distinguish instances (called pruning), and to improve predictive accuracy by reducing overfitting.

3.2. Random Forest Method

Random forests (RF) are an ensemble learning method which is developed based on aggregation of a large number of trees. RFs train a multitude of decision/regression trees and those decisions/regression trees vote for the mode of the classes (for classification) or the averaging (for regression), that is, for a tree based random forest [22]

$$RF(x) = sgn(\sum_{i=1}^{k} RT_i(x)) \quad (4).$$

Figure 2 is an illustration of the tree growing and forest building processes [22].

Suppose we have the training dataset c=(C_1,C_2…C_n) with Ci=(x_i,y_i) and the independent test case C_0 with predictor x_0.

(1) Sample the training set C with replacement to generate bootstrap resamples B_1, . . . , B_M.

(2) For each resample B_m, m = 1, . . . , M, grow a classification or regression tree T_m as described in section 3, except for the following modifications.
   a. At each split, only predictors in a randomly selected subset of predictors are considered as discussed in Section 4.2. Let p denote the total number of predictor variables in C.
   b. Each tree is grown until all nodes contain observations no more than the maximal terminal node size (MTN), a pre-specified parameter.

(3) For predicting the test case C_0 with covariate x_0, the predicted value by the whole RF is obtained by combining the results given by individual trees.

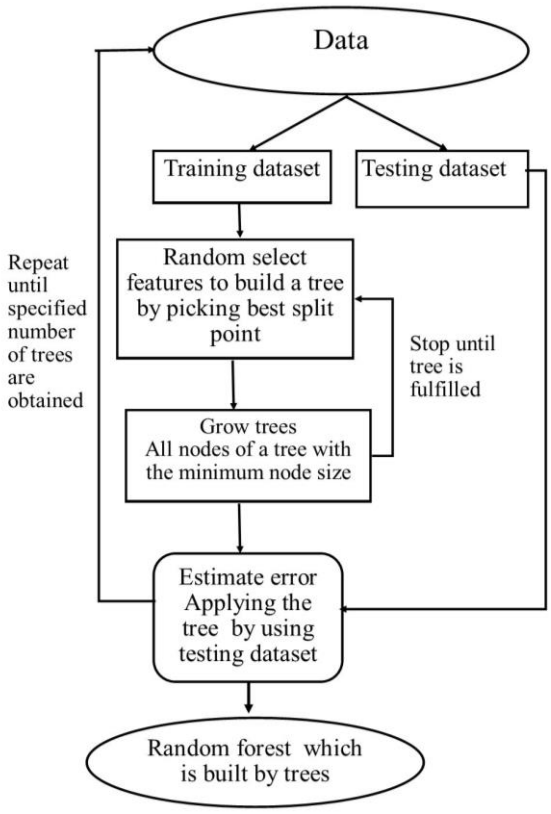


Figure 2. Flowchart of Random Forest Model Building.

RFs help reduce decision trees' probability of overfitting the training set; and have been introduced in the field of electricity load forecasting.

3.3. Model Transferability

We evaluate the transferability of models in four different ways:

(1) Use data from any arbitrary temperature zone for testing the models developed using data from the remaining four temperature zones;

(2) Evaluate the transferability of the models developed for one type of humidity zone to another humidity zone type;

(3) Similar to (2) but evaluate the transferability from one humidity zone type to another, for each of the temperature zones (3, 4, 5) which encompasses different humidity zone types;

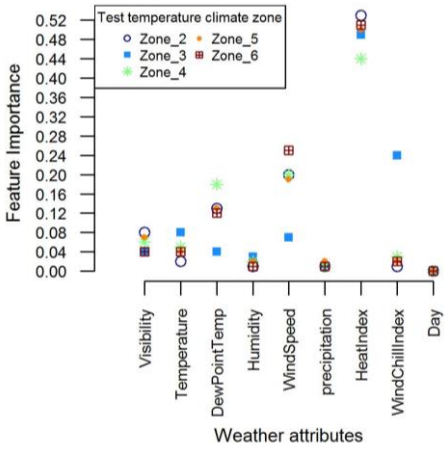
(4) similar to (1) but evaluate the transferability to one temperate zone from others, for each of the humidity zone type.

4. Results and Discussion

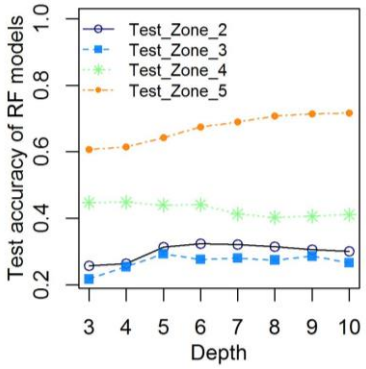
4.1. Transferability cross Temperature zones

WECC regions are corresponding to five different temperature zones (2, 3, 4, 5 and 6). The transferability of the RF load profile models across these temperature zones are evaluated by grouping four temperature zones for training, and the remaining one temperature zone is for testing. For example, if the training temperature zones are zones 3, 4, 5 and 6, then zone 2 will be the testing zone. Figure 3A shows the feature importance of five different RF models. Heat index is found to be important in each RF model. Wind chill index is important in the RF model for testing zones 2, 4, 5, and 6. Beside heat index, wind speed, dew point temperature, temperature and visibility all have different relative importance in the RF models.

Figure 3B shows the testing accuracy score of the RF models. For example, when the temperature zone 5 is considered as the testing zone, the finalized depth of the RF model is 7 and the testing accuracy is 0.62, which means that the model developed at other zones are transferrable. However, the testing accuracies for testing zones 2 and 3 are not high, as temperature zone 2 is much hotter during the summer and therefore corresponding to a different temperature-driven cooling mechanism, while temperature zone 3 has distinctly lower and narrower range of temperature.

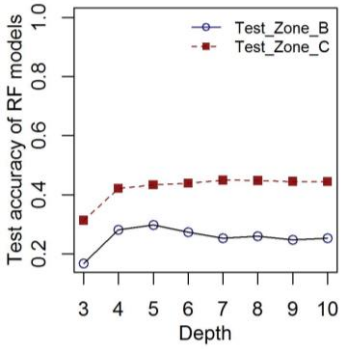


A. Feature importance



B. Test score

Figure 3. Results of Random forest models of different temperature zones.

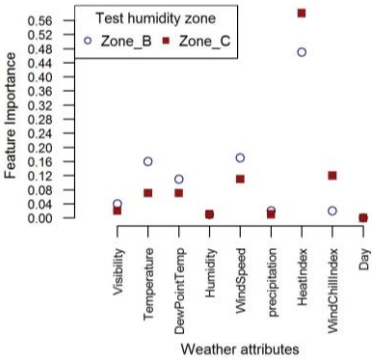


B. Test score

Figure 4. Random forest feature importance and test score of cooling of different DOE climate zones.

4.2 Transferability across humidity zones

Figure 4A shows feature importance in the RF models for different humidity zones (climate regimes A, B, C). When one humidity zone is used as the testing zone, the other zone is considered as training zones. The heat index is more important than other features in both the two models for zones B and C. Overall the importance of the nine features are comparable in these two models. The testing accuracy of these two models (see Figure 4B) are between 0.3 and 0.5, which indicates relatively weak transferability of the developed RF models for predicting load profiles, although such a level of accuracy is still acceptable when there is no other load profile to start with for power system planning and operations.

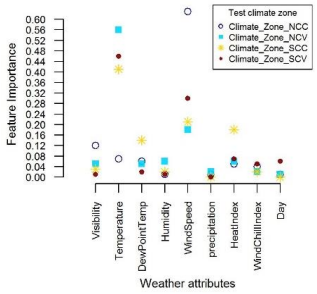


A. Feature importance

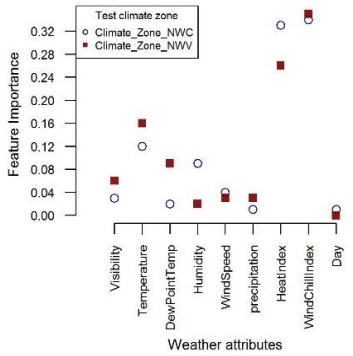
4.3. Transferability within temperature zones

The above analyses are done to evaluate the RF model transferability across climate zones, next we evaluate the transferability across regions but within each climate zone.

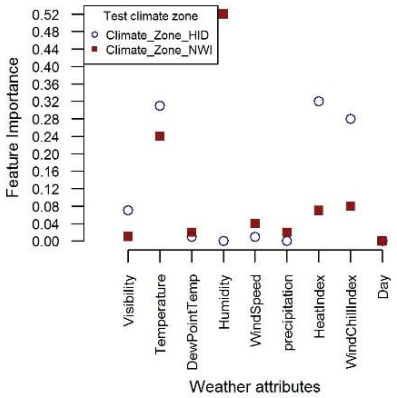
Figure 5 shows the feature importance of different RF models with training and testing data in the same temperature zones. For temperature zone 3 and 5, temperature is more important than other features. While for temperature zone 4, heat index and wind chill index play more important roles than other features. Generally, temperature is important to all the climate zones, while different climate zones have their own dominant features.



A. Feature importance for temperature zone 3



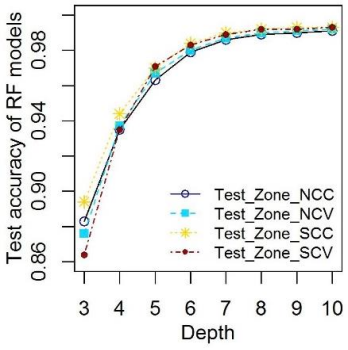
B. Feature importance for temperature zone 4



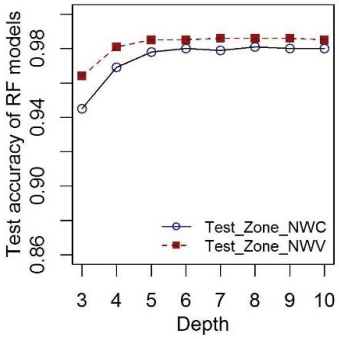
C. Feature importance for temperature zone 5

Figure 5. Feature importance in the RF models for each of the temperature zones.

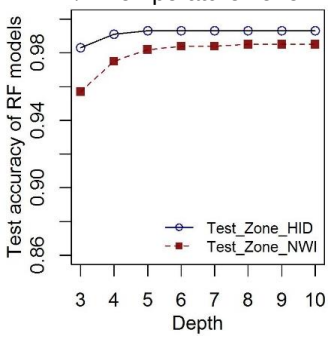
Figure 6 is the testing accuracy of the RF models developed for each of the three temperature zones 3, 4, and 5. The testing accuracy is about 90% and above when the RF model depth is 3 and larger. This means that in the same temperature zone the RF models are directly transferable. With further increase of depths, the increase of test accuracy is limited, which indicates that simple RF models may be sufficient for the three different temperature zones.



A. Temperature zone 3



B. Temperature zone 4

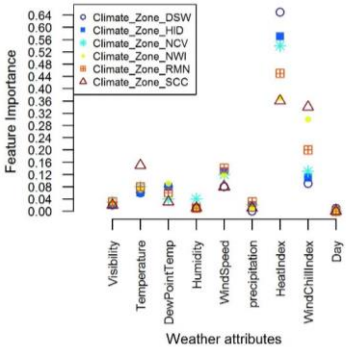


C. Temperature zone 5

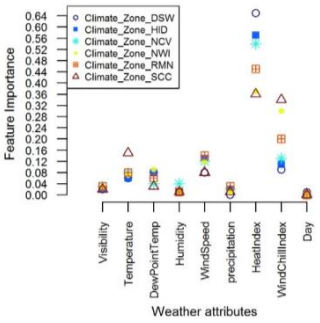
Figure 6. Random forest testing accuracy within temperature zones.

4.4. Transferability within humidity zones

Figures 7 and 8 are the feature importance plots and random forest test accuracy for the RF models developed within each of the two humidity zones (regimes B and C). Heat index is more important than other features for cooling in general. The testing accuracies are high; with a depth of 4 or larger, the models can achieve an accuracy higher than 0.90. When the depth is greater than six, the increase of the accuracy scores is little. Overall the transferability of RF models within the same humidity zones are high.

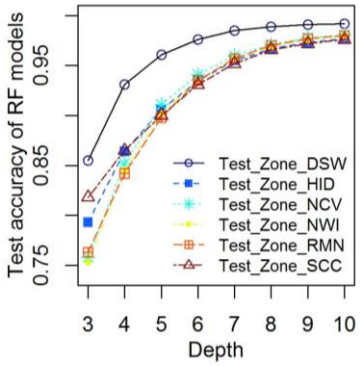


A. Humidity zone B

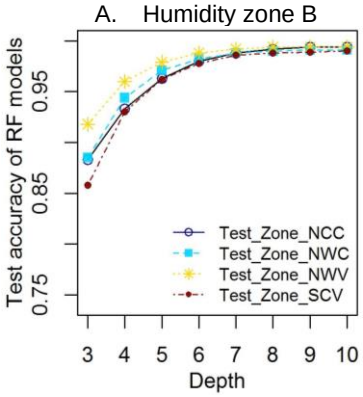


B. Humidity zone C

Figure 7. Feature importance in the RF models for each of the humidity zones.



A. Humidity zone B



B. Humidity zone C

Figure 8. Random forest testing accuracy within humidity zones.

5. Summary and Conclusions

In this study, we conducted comprehensive evaluation of ML (i.e., RF) model transferability across and within climate zones for predicting load profiles based on adjacent weather information. The transferability cross temperature and humidity zones are generally weak, unless the testing zone is located somewhere near the geographic center of the training zones such that the weather-driven demand mechanisms are comparable. It is therefore difficult to use the existing RF models to predict cooling load profile in different temperature zones or humidity zones. Despite that the model coefficients are not directly transferable, the model structure (e.g., the dominant features) is transferrable and consistent with findings from the literature. It can therefore provide guidance on developing zonal ML models for load profile approximation.

On the other hand, the RF models can be used with confidence if the target area is located in the same temperate and/or humidity zones as the regions providing training data. When the RF model depth is five or above, the testing accuracies within temperature zone and humidity zones are over 0.94, indicating a high level of model transferability within temperature zones and humidity zones. In practice, one can identify the climate zone for the target region without load profile information and integrate the corresponding zone-specific ML model with local weather data.

6. Acknowledgment

This work is supported by the U.S. Department of Energy (DOE) Office of Electricity Delivery and Energy Reliability as part of Advanced Grid Research and Development Program. The authors gratefully thank Mr. Ali Ghassemian from DOE for his continuing support, help, and guidance.

Pacific Northwest National Laboratory is operated by Battelle for DOE under Contract DE-AC05-76RL01830.

7. References

[1] D. Chassin, Y. Zhang, P. Etingov, D. James, D. Hatley, H. Kirkham, J. Kueck, X. Li, Y. Huang, and C. Chen, "ARRA Interconnection Planning-Load Modeling Activities," *PNNL-24425*: Pacific Northwest National Laboratory, 2015.

[2] B. Lesieutre, R. Bravo, R. Yinger, D. Chassin, H. Huang, N. Lu, I. Hiskens, and G. Venkataramanan, *Final Project Report Load Modeling Transmission Research*, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2012.

[3] L. Pereira, D. Kosterev, P. Mackin, D. Davies, J. Undrill, and W. J. I. T. o. P. S. Zhu, "An interim dynamic induction motor model for stability studies in the WSCC," vol. 17, no. 4, pp. 1108-1115, 2002.

[4] P. F. Village, "System Disturbances," 2002.

[5] Y. Liu, P. V. Etingov, S. Kundu, Z. Hou, Q. Huang, H. Zhou, M. Ghosal, D. P. James, J. Zhang, and Y. Xie, *Open-Source High-Fidelity Aggregate Composite Load Models of Emerging Load Behaviors for Large-Sale Analysis*, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2020.

[6] Y. Liu, Z. Hou, P. Etingov, and H. Zhou, "Update of Residential Load Profile for WECC Load Composition Model Using Cross-Correlation Method." pp. 1-5.
[7] H. Zhou, Z. Hou, P. Etingov, and Y. Liu, "Machine-Learning-Based Investigation of the Associations Between Residential Power Consumption and Weather Conditions." pp. 85-91.
[8] P. Xu, J. Huang, R. Jin, and G. Yang, *Measured energy performance of a US-China demonstrationenergy-efficient office building*, Ernest Orlando Lawrence Berkeley NationalLaboratory, Berkeley, CA (US), 2006.
[9] J. W. Taylor, and R. J. I. J. o. F. Buizza, "Using weather ensemble predictions in electricity demand forecasting," vol. 19, no. 1, pp. 57-70, 2003.
[10] K. K. Wan, D. H. Li, D. Liu, J. C. J. B. Lam, and Environment, "Future trends of building heating and cooling loads and energy consumption in different climates," vol. 46, no. 1, pp. 223-234, 2011.
[11] R. Lindberg, A. Binamu, and M. Teikari, "Five-year data of measured weather, energy consumption, and time-dependent temperature variations within different exterior wall structures," *Energy and Buildings*, vol. 36, no. 6, pp. 495-501, 2004.
[12] M. Beccali, M. Cellura, V. Lo Brano, and A. Marvuglia, "Short-term prediction of household electricity consumption: Assessing weather sensitivity in a Mediterranean area," *Renewable and Sustainable Energy Reviews*, vol. 12, no. 8, pp. 2040-2065, 2008.
[13] M. Braun, H. Altan, and S. J. A. E. Beck, "Using regression analysis to predict the future energy consumption of a supermarket in the UK," vol. 130, pp. 305-313, 2014.
[14] B. Yildiz, J. I. Bilbao, A. B. J. R. Sproul, and S. E. Reviews, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," vol. 73, pp. 1104-1122, 2017.
[15] J. Huo, T. Shi, and J. Chang, "Comparison of Random Forest and SVM for electrical short-term load forecast with different data sources." pp. 1077-1080.
[16] G. Dudek, "Short-term load forecasting using random forests," *Intelligent Systems' 2014*, pp. 821-828: Springer, 2015.
[17] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. J. A. E. Mochida, "Applying support vector machine to predict hourly cooling load in the building," vol. 86, no. 10, pp. 2249-2256, 2009.
[18] H. Mori, and N. Kosemura, "Optimal regression tree based rule discovery for short-term load forecasting." pp. 421-426.
[19] R. S. Briggs, R. G. Lucas, and Z. T. J. A. T. Taylor, "Climate classification for building energy codes and standards: Part 2-Zone definitions, maps, and comparisons," vol. 109, pp. 122, 2003.
[20] R. S. Briggs, R. G. Lucas, and Z. T. J. A. T. Taylor, "Climate classification for building energy codes and standards: Part 1-development process," vol. 109, pp. 109, 2003.
[21] L. Breiman, *Classification and regression trees*: Routledge, 2017.
[22] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*: Springer series in statistics New York, 2001.